ENVIRONMETRICS
Environmetrics 2001; **12**: 451–467 (DOI: 10.1002/env.472)

Compositional receptor modeling

Dean Billheimer[*,†]

*National Research Center for Statistics and the Environment, Department of Statistics,
Box 354322, University of Washington, Seattle, WA 98195-4322, U.S.A*

SUMMARY

Receptor models apportion an ambient mixture of pollutants to the contributing pollution sources. Often, neither the number of sources nor their chemical profiles are known precisely. The dual goals of modeling are to estimate the chemical 'signature' of the sources, and to characterize the mixing process. The author develops a novel modeling approach for receptor data where all model components are compositions (i.e. vectors of proportions). This approach maintains positivity and summation constraints for source contributions and chemical profiles. Further, it incorporates available prior knowledge regarding the source chemical profiles. Including prior knowledge allows parameter estimation while avoiding restrictive assumptions regarding presence or absence of chemical tracers. This approach is illustrated by modeling air pollution data collected from a receptor near Juneau, Alaska. The compositional model produces point estimates of source profiles and mixing proportions similar to those obtained in a previous study. However, interval estimates for mixing proportions are roughly 30 per cent shorter than those found previously. Copyright © 2001 John Wiley & Sons, Ltd.

KEY WORDS: compositional data; receptor model; source apportionment; air pollution; convex mixture

1. INTRODUCTION TO SOURCE RECEPTOR MODELING

Air quality management is a difficult problem with important consequences for human and environmental health. The difficulties arise primarily from problems with pollution measurement and transport: identification of sources, estimation of emission rates, physical transport of substances, and physical and chemical transformation processes occurring during transport (Hopke, 1999). Source apportionment, or receptor, models address these issues by analyzing pollution concentrations measured in ambient air. These models aim to identify pollution sources and apportion pollutant loadings to those sources. Observations consist of a convex mixture of chemical species originating from different sources. In the most general case, neither the number of sources nor the individual source chemical profiles are known. The dual goals of receptor modeling are to estimate the chemical 'signature' of the sources, and to characterize the mixing process. A comprehensive review of receptor modeling as well as source-oriented dispersion models is available in Hopke (1991).

[*] Correspondence to: Dean Billheimer, National Research Center for Statistics and the Environment, Department of Statistics, Box-354322, University of Washington, Seattle, WA 98195–4322, U.S.A.
[†] E-mail: dean@stat.washington.edu

Received 5 July 2000
Revised 26 September 2000

To illustrate the air pollution receptor problem, consider the study described in Aldershof and Ruppert (1987) and Bandeen-Roche (1994, hereafter referenced as BR). Researchers collected $n = 50$ ambient air samples at a receptor near Juneau, Alaska. Each observation is a 'daily' (time averaged) vector of the relative mass of five chemical species (fluoranthene, benzoanthracene, chrysene, benzofluoranthene, and pyrene). Two sources are believed to contribute to local pollution: wood-stove smoke and motor vehicle emissions. Although much is known about the chemical profiles of these sources, they are not known precisely. Further, little is known about the mixing process. The study's goal is to estimate the contribution of wood-stoves to the local pollution load. Estimation of the individual source profiles is of secondary interest.

In this article I develop an approach to receptor modeling that incorporates prior knowledge of pollution sources to estimate source profiles. Further, I use a novel statistical error structure, based on principles of compositional data, that ensures model parameter estimates conform to physically based constraints. I maintain the Juneau receptor example through the remainder of the paper to illustrate this modeling approach.

Current receptor models are based on the principle of chemical mass balance. That is, the total amount of a chemical species present in a sample is the sum of the contributions of the individual sources. For a fixed number of sources, p, observation i $(i = 1, 2, \ldots, n)$ is modeled as a linear combination of sources' chemical species,

$$\mathbb{E}[\mathbf{Y}_i] = \sum_{j=1}^{p} \alpha_{ji}\boldsymbol{\theta}_j = [\boldsymbol{\theta}_1 \mid \boldsymbol{\theta}_2 \mid \ldots \mid \boldsymbol{\theta}_p] \begin{bmatrix} \alpha_{1i} \\ \alpha_{2i} \\ \vdots \\ \alpha_{pi} \end{bmatrix} = \boldsymbol{\Theta}\boldsymbol{\alpha}_i \tag{1}$$

Here, $\mathbf{Y}_i$ is a vector of concentrations of k chemical species, $\boldsymbol{\alpha}_i$ is a p-vector of mixing coefficients, and $\boldsymbol{\theta}_j$ (for $j = 1, 2, \ldots, p$) is a k-vector describing the chemical profile for source j. Often some form of measurement error is included in the model.

In the case of the Juneau study, this model reduces to

$$\mathbb{E}[\mathbf{Y}_i] = \alpha_{wi}\boldsymbol{\theta}_w + \alpha_{mi}\boldsymbol{\theta}_m \tag{2}$$

Here, α_{wi} and α_{mi} are the contributions of wood-smoke and motor vehicle exhaust, respectively, with source profiles $\boldsymbol{\theta}_w$ and $\boldsymbol{\theta}_m$. In this case $\mathbf{Y}_i, \boldsymbol{\theta}_w$ and $\boldsymbol{\theta}_m$ are vectors of the relative masses of the five measured chemical species. The quantity of primary interest is the amount of pollution attributable to wood-stoves (α_w). To determine this we must also estimate the source chemical profiles.

When data are measured as *relative* concentrations (as in the Juneau study), there are constraints on the quantities in Equation (1) not generally applicable to all source apportionment models. Observations, $\mathbf{Y}_i$, and the source profiles, $\boldsymbol{\theta}_j$, are assumed standardized to compositional form (all elements non-negative, and all elements sum to one). Further, $\alpha_{ji} > 0$, and $\sum_{j=1}^{p} \alpha_{ji} = 1$; thus $\boldsymbol{\alpha}_i$ is also a composition. There are advantages to modeling relative, rather than absolute, concentrations. Models for relative concentrations are more widely applicable since some aggregates are measured only in compositional form (whereas 'raw' concentrations can always be normalized to compositions). More importantly, relative concentrations may be measured with greater precision than absolute concentrations (BR; Kowalczyk *et al.*, 1978). However, requiring observations to be compositions limits inference if the total amount of pollutants collected is correlated with a particular subset of the sources. This problem might be addressed by incorporating the total as a predictive covariate (BR).

1.1. *Modeling difficulties*

As noted by BR, Park *et al.* (1999), and others, a number of difficulties arise in fitting Equation (1). Because $\boldsymbol{\alpha}_i$ is a vector of mixing proportions (for day *i*) its elements must satisfy positivity and summation constraints. Such constraints can be awkward to include in parameter estimation. A second difficulty is encountered when the $\boldsymbol{\theta}_j$s are assumed fixed and each observation, Y_i, is associated with a unique $\boldsymbol{\alpha}_i$. This is an example of the incidental parameter problem identified by Neyman and Scott (1948) and addressed in detail by Kiefer and Wolfowitz (1956). In related problems the fixed $\boldsymbol{\theta}_j$s are often the quantities of interest, and the incidental parameter(s) are treated as a 'nuisance' in estimation. However, for receptor modeling problems in which the mixing process is of interest, the incidental parameters are paramount since they describe the amount of pollution attributable to different sources.

Finally, if the source profiles ($\boldsymbol{\theta}_j$) are unknown, the parameters of Equation (1) are not identifiable. Other researchers (e.g. BR; Park *et al.*, 1999) have thoroughly examined identifiability conditions. Their approaches parallel considerations described in Reiersol (1950) and Lindsay (1983), and require each source to be absent from at least one observation. Alternatively, the presence or absence of 'tracer' elements (chemical species known to be absent from a source, or confined to a single source) can also indicate the presence/absence of a source. Modern approaches to receptor modeling (see, for example, Park *et al.*, 1999) require $p-1$ species to be absent from each source. When such conditions are met, the 'source polytope' can be defined and the model parameters identified (BR).

In the remainder of this article, I propose a modeling approach for a restricted version of the source apportionment problem. The problem characteristics are as follows:

1. observations are the relative concentrations of the chemical species under study
2. the number of sources, *p*, is known
3. partial information is available about the individual source chemical profiles.

To accommodate compositional observations, I use a non-additive error structure described in Billheimer *et al.* (1997, 2000), and based on Aitchison's (1986) perturbation operator. This structure ensures that compositional quantities satisfy positivity and summation constraints. The incidental parameter problem is alleviated by modeling the mixing proportions as (unobservable) realizations from a distribution (Kiefer and Wolfowitz, 1956). This modeling choice also aids interpretation of the mixing process by shifting our focus to characteristics of the distribution. Finally, information about source profiles is incorporated into the modeling framework by means of informative (hyper) prior distributions. Prior information about sources avoids the identifiability concerns mentioned above, and allows estimation of model parameters. No assumptions regarding the observations or missing species are required.

The next section describes the general model formulation and methods for inference,while Section 3 specifies the implementation for analysis of the Juneau, AK, receptor data. Section 4 presents the results of this analysis, and compares these results with BR. The final section compares the compositional model with other recent approaches, and describes directions for future development.

2. MODEL FORMULATION

A statistical source apportionment model for relative concentrations is formulated as follows:

$$\mathbf{Y}_i = \boldsymbol{\Theta}\,\boldsymbol{\alpha}_i \oplus \boldsymbol{\epsilon}_i \tag{3}$$

where $\mathbf{Y}_i$, $\boldsymbol{\Theta}$ and $\boldsymbol{\alpha}_i$ are interpreted as before, but are restricted to be compositional quantities. Note that $\mathbf{Y}_i$ is an element of the $(k-1)$-dimensional simplex (∇^{k-1}). That is, $Y_{it} > 0 \quad \forall t$, and $\sum_{t=1}^{k} Y_{it} = 1$. Further, the columns of $\boldsymbol{\Theta}$ are elements of ∇^{k-1}, and the $\boldsymbol{\alpha}_i$s are elements of ∇^{p-1}. Multivariate random error, $\boldsymbol{\epsilon}_i \in \nabla^{k-1}(i = 1, 2, \ldots, n)$, is assumed independent and identically distributed from a logistic normal distribution with location parameter vector $\mathbf{0}_{k-1} = (0, \ldots, 0)'(k-1\text{-vector})$ and dispersion matrix $\sum_{\boldsymbol{\epsilon}}$.

As the notation suggests, the symbol '$\oplus$' denotes an addition operator for compositional quantities, where addition is defined on ∇^{k-1}(see Appendix 4.1 for details). This operation, Aitchison's (1986) perturbation operator, provides a natural definition for 'additive error' for compositional data (Billheimer *et al.*, 1997, 2000). Briefly, for two compositional quantities $\mathbf{u}$ and $\mathbf{v}$ in ∇^{k-1},

$$\mathbf{u} \oplus \mathbf{v} = \left(\frac{u_1 v_1}{\sum_{i=1}^{k} u_i v_i}, \frac{u_2 v_2}{\sum_{i=1}^{k} u_i v_i}, \ldots, \frac{u_k v_k}{\sum_{i=1}^{k} u_i v_i} \right)' \tag{4}$$

resulting in another composition in ∇^{k-1}. Billheimer *et al.* (2000) show that the perturbation operator can be used to construct an algebra and a complete, normed, vector space for compositions. This construction allows the usual notions of additive error and projection (estimation) to be extended to compositional data.

To complete the model specification, I cast $\boldsymbol{\theta}_j$ and $\boldsymbol{\alpha}_i$ in a hierarchical framework. Each source profile, $\boldsymbol{\theta}_j$, is equipped with an informative prior distribution describing (partial) knowledge of the relative concentrations of its chemical species. Typically, a logistic normal is used for this specification, although any compositional distribution suffices. Less information is available about the mixing proportions. The $\boldsymbol{\alpha}_i$s are modeled as independent draws from a logistic normal distribution with unknown location parameters $\boldsymbol{\mu}_\alpha (\in \Re^{p-1})$ and dispersion matrix Γ. These parameters are given diffuse, but proper, conjugate hyper-prior distributions. I use Markov chain Monte Carlo (MCMC) for inference about model parameters. While a convex combination of compositional quantities presents difficulties for analytic description (Aitchison and Bacon-Shone, 1999), it is easily accommodated by the MCMC algorithm.

To summarize the distributional assumptions,

$$\pi(\boldsymbol{\Theta}, \boldsymbol{\alpha}_i, \boldsymbol{\epsilon}_i, \boldsymbol{\mu}_\alpha, \boldsymbol{\Gamma}, \boldsymbol{\Sigma}_\epsilon) = \pi(\boldsymbol{\alpha}_i \mid \boldsymbol{\mu}_\alpha, \boldsymbol{\Gamma})\, \pi(\boldsymbol{\epsilon}_i \mid \boldsymbol{\Sigma}_\epsilon)\, \pi(\boldsymbol{\mu}_\alpha)\, \pi(\boldsymbol{\Gamma})\, \pi(\boldsymbol{\Sigma}_\epsilon)\, \pi(\boldsymbol{\Theta})$$

where

$$\begin{aligned} &\boldsymbol{\epsilon}_i \sim L^{k-1}(\mathbf{0}_{k-1}, \boldsymbol{\Sigma}_\epsilon); \quad \boldsymbol{\Sigma}_\epsilon^{-1} \sim \text{Wishart}(a\, \mathcal{N}, \rho) \\ &\boldsymbol{\theta}_j \sim L^{k-1}(\mu_{\boldsymbol{\theta}_j}, \Sigma_{\boldsymbol{\theta}_j}) \\ &\quad \boldsymbol{\alpha}_i \sim L^{p-1}(\boldsymbol{\mu}_\alpha, \boldsymbol{\Gamma}); \quad \boldsymbol{\mu}_\alpha \sim N_{p-1}(\boldsymbol{\eta}, \Psi); \quad \boldsymbol{\Gamma}^{-1} \sim \text{Wishart}\,(b\, \mathcal{N}, \delta) \end{aligned}$$

Here, $L^{k-1}(\boldsymbol{\mu}, \Sigma)$ denotes the logistic normal distribution of dimension $k-1$ with location parameter vector $\boldsymbol{\mu}$ and dispersion matrix Σ. Specification of the parameter values is postponed to the next section.

For notational convenience, let $\phi(.)$ denote Aitchison's (1986) additive log-ratio transformation. That is, for $z \in \nabla^{k-1}$

$$\phi(z) = \left[\log\left(\frac{z_1}{z_k}\right), \log\left(\frac{z_2}{z_k}\right), \ldots, \log\left(\frac{z_{k-1}}{z_k}\right)\right]' \tag{5}$$

Thus, $\phi(.)$ is a bijection mapping ∇^{k-1} onto $\Re^{k-1}$.

Combining likelihood and priors, the joint posterior distribution is proportional to the following expression:

$$\begin{aligned}
&\pi(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \ldots, \boldsymbol{\theta}_p, \boldsymbol{\alpha}_1, \boldsymbol{\alpha}_2, \ldots, \boldsymbol{\alpha}_n, \boldsymbol{\mu}_\alpha, \boldsymbol{\Gamma}, \boldsymbol{\Sigma}_\epsilon \mid \boldsymbol{y}) \propto \\
&\prod_{i=1}^{n}\left\{ \mid \boldsymbol{\Sigma}_\epsilon \mid^{-1/2} \prod_{t=1}^{k} [\boldsymbol{y}_i]_t^{-1} \exp\left\{-\frac{1}{2}[\phi(\boldsymbol{y}_i) - \phi(\boldsymbol{\Theta\alpha}_i)]' \, \boldsymbol{\Sigma}_\epsilon^{-1}[\phi(\boldsymbol{y}_i) - \phi(\boldsymbol{\Theta\alpha}_i)]\right\}\right. \\
&\left.\times \prod_{j=1}^{p} \mid \boldsymbol{\Gamma} \mid^{-1/2} \alpha_{ij}^{-1} \exp\left\{-\frac{1}{2}[\phi(\boldsymbol{\alpha}_i) - \boldsymbol{\mu}_\alpha]' \, \boldsymbol{\Gamma}^{-1}[\phi(\boldsymbol{\alpha}_i) - \boldsymbol{\mu}_\alpha]\right\}\right\} \\
&\times \exp\left\{-\frac{1}{2}(\boldsymbol{\mu}_\alpha - \boldsymbol{\eta})'\Psi^{-1}(\boldsymbol{\mu}_\alpha - \boldsymbol{\eta})\right\} \prod_{j=1}^{p} \exp\left\{-\frac{1}{2}\left[\phi(\boldsymbol{\theta}_j) - \boldsymbol{\mu}_{\boldsymbol{\theta}_j}\right]' \boldsymbol{\Sigma}_{\boldsymbol{\theta}_j}^{-1}\left[\phi(\boldsymbol{\theta}_j) - \boldsymbol{\mu}_{\boldsymbol{\theta}_j}\right]\right\} \times \pi(\boldsymbol{\Gamma}) \times \pi(\boldsymbol{\Sigma}_\epsilon)
\end{aligned}$$

where $[\boldsymbol{y}_i]_t$denotes the tth element of $\boldsymbol{y}_i$.

2.1. *Markov chain Monte Carlo estimation*

From this expression, full conditional distributions for $\boldsymbol{\theta}_j, \boldsymbol{\alpha}_i, \boldsymbol{\mu}_\alpha, \boldsymbol{\Gamma}$ and, $\boldsymbol{\Sigma}_\epsilon$ can be obtained (up to normalizing constants) for use in MCMC sampling (Besag and Green, 1993). The choice of conjugate distributions results in a multivariate normal distribution for $\boldsymbol{\mu}_\alpha$ and inverse Wishart distributions for $\boldsymbol{\Gamma}$ and $\boldsymbol{\Sigma}_\epsilon$. These components can be updated by direct (Gibbs) sampling. Conversely, $\boldsymbol{\Theta}$ and $\boldsymbol{\alpha}_i$ do not have simple analytic distributional forms. The Metropolis–Hastings algorithm can be used to update these components.

3. IMPLEMENTATION FOR JUNEAU AIR POLLUTION DATA

I apply the model formulated above to analyze data from the Juneau, AK, air pollution receptor. First, I select prior distributions for the wood-smoke and motor vehicle emission profiles to facilitate comparison with BR's analysis of these data. This is referenced as 'model A' in the following sections. Then I re-run the analysis using less restrictive prior information by considering both source profiles to be unknown (referenced as 'model B'). Note that this is not an exhaustive analysis of the Juneau receptor data. The focus is to compare results with previous modeling efforts, and to evaluate the sensitivity of results to restrictions on prior distributions.

Recall that $n = 50$ daily ambient air samples were collected, and that the data were comprised of the relative mass of five chemical species: fluoranthene, benzoanthracene (benzo a), chrysene, benzofluoranthene (benzo b), and pyrene. Two pollution sources were believed to contribute to the pollution load: wood-stove smoke, and motor vehicle emissions. BR's analysis assumes that the chemical profile for wood-stove smoke is ‘known’, but that motor vehicle exhaust profile is unknown. Both are considered constant. The goal is to estimate the contribution due to wood-smoke. Achieving this goal requires estimation of the motor vehicle exhaust profile.

I make similar assumptions in specifying prior distributions, namely:

- Wood-smoke source composition is assumed known and fixed at the concentrations used by BR.
- The motor vehicle emission composition prior distribution is centered at BR’s MLE. The prior variance of this distribution is specified to be informative, but to retain substantial variability for $\boldsymbol{\theta}_m$.
- Daily mixing proportions are modeled as independent, identically distributed.
- Prior distributions for mixing parameters and error variance are quite diffuse (but proper).

3.1. *Model A specification*

The prior distributions are defined as follows:

$$\boldsymbol{\theta}_m \sim L^{k-1}(\boldsymbol{\eta}_m, a\,\mathcal{N})$$

The logistic normal location parameter vector, $\boldsymbol{\eta}_m$, is more easily interpreted as a composition, thus, $\phi^{-1}(\boldsymbol{\eta}) = (0.04, 0.08, 0.29, 0.35, 0.24)'$. This distribution ‘center’ corresponds to the maximum likelihood estimate from BR’s analysis. The dispersion matrix, $\mathcal{N}$, specifies a ‘null’ correlation structure between log-ratio transformed compositions. That is, a priori one may consider the compositional elements ‘independent except for the summation constraint’ (Aitchison, 1986; Billheimer *et al.*, 1997). This matrix has the form

$$\mathcal{N} = \begin{bmatrix} 2 & 1 & 1 & 1 \\ 1 & 2 & 1 & 1 \\ 1 & 1 & 2 & 1 \\ 1 & 1 & 1 & 2 \end{bmatrix}$$

The value $a = 0.10$ is chosen for the motor vehicle profile, and provides substantial variability for the elements of $\boldsymbol{\theta}_m$. Table 1 summarizes medians and marginal 95 per cent prior probability intervals for each chemical species of the motor vehicle emission profile.

The mixing proportions $\boldsymbol{\alpha}_i$ are independently sampled a priori from a logistic normal distribution with location parameter $\boldsymbol{\mu}_\alpha$ and dispersion matrix $\boldsymbol{\Gamma}$. Because $p = 2$ sources, $\boldsymbol{\mu}_\alpha$ and $\boldsymbol{\Gamma}$ are both scalar

Table 1. 95% prior probability intervals for motor vehicle emissions

	Quantile		
Chemical species	0.025	0.50	0.975
Fluoranthene	0.019	0.040	0.076
Pyrene	0.039	0.079	0.147
Benzo (a)	0.160	0.293	0.451
Chrysene	0.201	0.344	0.519
Benzo (b)	0.129	0.244	0.388

quantities. I center the prior distribution for $\boldsymbol{\mu}_\alpha$ at zero, with a variance 0.8. This corresponds a 95% prior probability interval for $\phi^{-1}(\boldsymbol{\mu}_\alpha)$ ranging from approximately $(0.1, 0.9)'$ to (0.9, 0.1)', with the mode at (0.5, 0.5)'. The prior variance for $\boldsymbol{\alpha}_i$ is an inverse gamma distribution with shape parameter 1.0 and scale parameter 100. This specifies a diffuse but proper prior for $\boldsymbol{\Gamma}$.

Finally, the measurement error dispersion matrix, $\boldsymbol{\Sigma}_\epsilon$, is modeled with an inverse Wishart distribution with parameters $b\mathcal{N}$ and ρ degrees of freedom. For this problem, $b = 0.1$ and $\rho = 4 = k - 1$. Again, this specifies a diffuse, proper prior for the measurement error.

3.2. *Markov chain Monte Carlo sampling*

MCMC is used to sample the posterior distribution for $\boldsymbol{\theta}_m, \boldsymbol{\alpha}_i, \boldsymbol{\mu}_\alpha, \boldsymbol{\Gamma}$, and $\boldsymbol{\Sigma}_\epsilon$. Quantities used in the convex combination, $\boldsymbol{\theta}_m$ and $\boldsymbol{\alpha}_i$, are updated using the Metropolis–Hastings algorithm, while $\boldsymbol{\mu}_\alpha, \boldsymbol{\Gamma}$, and $\boldsymbol{\Sigma}_\epsilon$ are updated are updated using Gibbs sampling. Each parameter was updated once per sampling cycle, and the chain sampled for 100 000 cycles (after a burn-in period of 1000 cycles). Realizations from every 20th cycle are saved for further analysis. This subsampling is done solely to reduce storage space requirements and post-MCMC processing. Results are not sensitive to starting states of the chain, and appear to converge quickly to the limit distribution. Visual inspection of sampler output suggests that mixing is somewhat slower than in better identified hierarchical models. However, comparison of multiple chains initiated at different starting points, and sampler diagnostics (gibbsit, Raftery and Lewis, 1992), indicate run length to be adequate for reliable inference.

3.3. *Model B specification*

Next, I relax the assumption that the wood-smoke emission profile is known, and model it via an informative prior distribution. I center the prior distribution at the 'fixed' value in the previous analysis (used by BR), and specify a prior variance identical to that of the motor vehicle emission profile (e.g. 0.1 $\mathcal{N}$). The 0.025, 0.50, and 0.975 quantiles of the prior distribution for wood-smoke emissions are presented in Table 2.

The Juneau data are re-analyzed with this prior distribution for wood-smoke emissions. All other assumptions and distributions are identical to those described in model A. The results from this model are summarized in the next section.

Table 2. 95% prior probability intervals for wood-smoke emissions

	Quantile		
Chemical species	0.025	0.500	0.975
Fluoranthene	0.201	0.346	0.509
Pyrene	0.156	0.280	0.434
Benzo (a)	0.052	0.103	0.183
Chrysene	0.057	0.112	0.198
Benzo (b)	0.082	0.159	0.270

4. MODELING RESULTS FOR JUNEAU DATA

4.1. *Model A results*

Using the model A specification to fit the Juneau receptor data results in a sample from the joint posterior distribution of $\boldsymbol{\theta}_m, \boldsymbol{\alpha}_i, \boldsymbol{\mu}_\alpha, \boldsymbol{\Gamma}$, and $\boldsymbol{\Sigma}_\epsilon$. Figure 1 shows point estimates and credible intervals for daily proportion attributable to wood-smoke.

Points indicate median daily proportion estimates, while pointwise 50 and 95 per cent credible intervals are indicated by dashed and dotted lines, respectively. The plot shows that daily 95 per cent credible intervals range from roughly 0.2 to 0.7. This wide range of values suggests that we have little information for estimating daily mixing proportions. Focusing on the median (point) estimates, we also see substantial day-to-day variability in the proportion attributable to wood-smoke (range 0.33 to 0.52). Further, the observation from day 41 and possibly days 3 and 28 are suggestive of outlying observations. Finally, there is slight evidence of an upward trend in wood-smoke proportion through the time series. Regressing the log-ratio transformed point estimates against day reveals a positive slope (0.003), with a marginally significant p-value (0.035). (NB. This trend disappears if days 3 and 41 are omitted.) These last remarks are not intended as analysis 'results'. Instead, they are included to indicate that better, formal methods for diagnostics and inclusion of covariates are needed for analysis of receptor data.

The marginal distribution for the proportion attributable to wood-smoke is summarized by the histogram in Figure 2. The estimated (median) contribution of wood-smoke is 0.41 (denoted by the '$+$'), and an approximate 95 per cent credible interval (endpoints denoted by '*') is (0.27, 0.60).

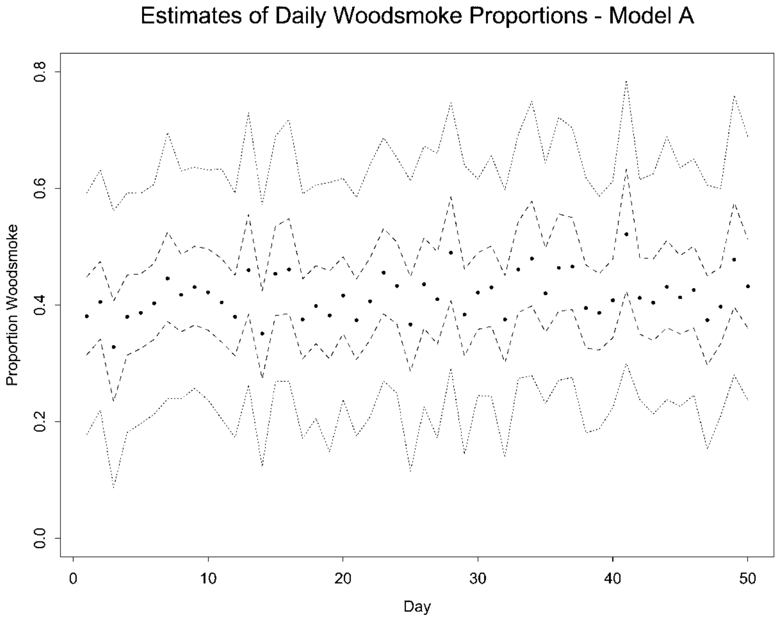

Figure 1. Point estimates and credible intervals for daily wood-smoke proportion. Points indicate median proportion attributable to wood-smoke for each daily observation. 50% and 95% credible intervals are indicated by dashed and dotted lines, respectively

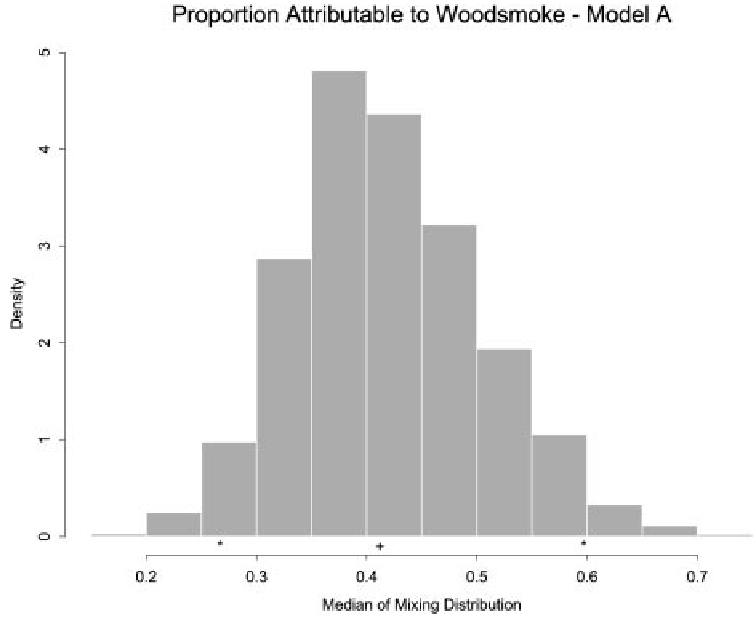

Figure 2. Histogram for median mixing proportions. This histogram shows the distribution of MCMC realizations for the median of the mixing proportion distribution, $\boldsymbol{\mu}_\alpha$. The median of this distribution is approximately 0.41, and a 95% credible interval is (0.27, 0.60)

These values compare favorably with those obtained by BR. She obtained a maximum likelihood estimate (including outliers) of 0.37 with a 95 per cent confidence interval of (0.10, 0.56). Note that the credible interval above is approximately 30 per cent narrower than BR's confidence interval. (One should also note that with two outliers removed, BR's point estimate is 0.42, with a much narrower confidence interval of 0.34 to 0.49.) Factors contributing to this narrower interval are discussed in the next section.

Finally, the estimated motor vehicle emission profile was similar to that obtained by BR, and is summarized in Table 3. The similarity is not surprising since, to facilitate comparison, the prior location parameter vector for $\boldsymbol{\theta}_m$ was set at BR's maximum likelihood estimate.

Slight discrepancies between estimates exist for chrysene and benzo (b). The hierarchical compositional model estimates slightly more chrysene and less benzo (b) than does the method of

Table 3. Motor vehicle emission profile

Chemical species	Estimate	MLE from BR
Fluoranthene	0.039	0.040
Pyrene	0.075	0.079
Benzo (a)	0.294	0.293
Chrysene	0.395	0.344
Benzo (b)	0.197	0.244

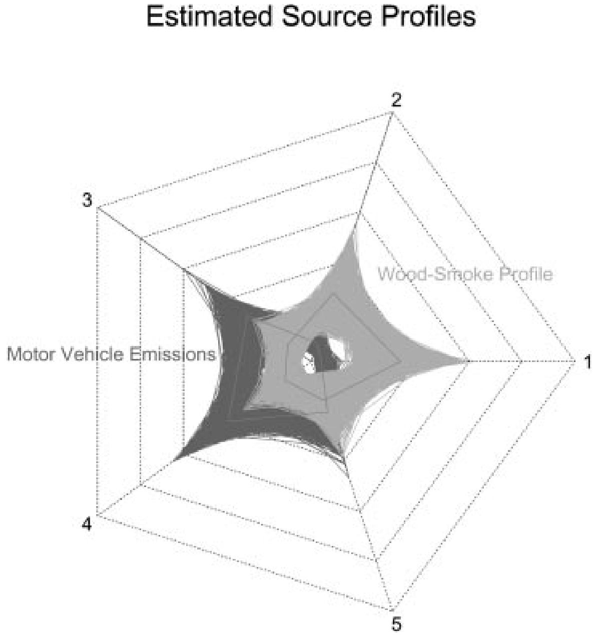

Figure 3. Sample of posterior distribution of source profiles. The webplot shows sample from the posterior distribution of wood-smoke and motor vehicle emission profiles. Dark lines superimposed on the figure represent point estimate compositions for each source

BR. However, marginal (elementwise) 95 per cent credible intervals for chrysene (0.27, 0.53) and for benzo (b) (0.11, 0.31) are clearly consistent with BR's estimates.

4.2. *Model B results*

MCMC sampling from model B again provides a sample from the joint posterior distribution. However, with this model both the wood-smoke source profile and motor vehicle emission profile are included in the posterior distribution. Figure 3 shows posterior estimates from these distributions.

The posterior distributions for both wood-smoke and motor vehicles are very similar to their prior distributions. This similarity likely reflects (i) that the prior distributions are centered near the maximum likelihood estimators, and/or (ii) there is little information in the data for updating the prior distribution. This issue is discussed further in the next section.

Figure 4 shows point estimates and credible intervals for daily proportion attributable to wood-smoke. Clearly, there are strong similarities between Figures 4 and 1. One difference of note is that the variability in Figure 4 is smaller. Credible intervals for daily proportions are about 0.10 narrower for model B than in model A. Indeed, point estimates also shrink toward the central value of 0.42. Increasing the source profile uncertainty appears to reduce observed variability in estimation of the mixing parameters.

Figure 5 summarizes the posterior distribution for the 'mean' mixing proportion, $\boldsymbol{\mu}_\alpha$. Again, results are similar to those obtained from model A; the primary difference is reduced variability in the present model. The median for this distribution is about 0.42 with a 95 per cent credible interval of (0.30,

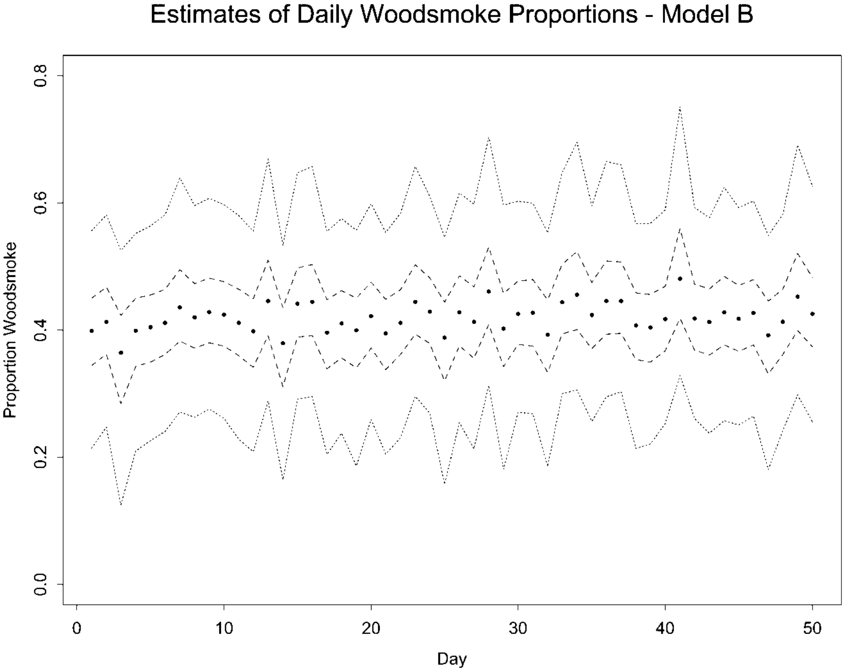

Figure 4. Point estimates and credible intervals for daily wood–smoke proportion – model B. Points indicate median proportion attributable to wood-smoke for each daily observation. 50% and 95% credible intervals are indicated by dashed and dotted lines, respectively

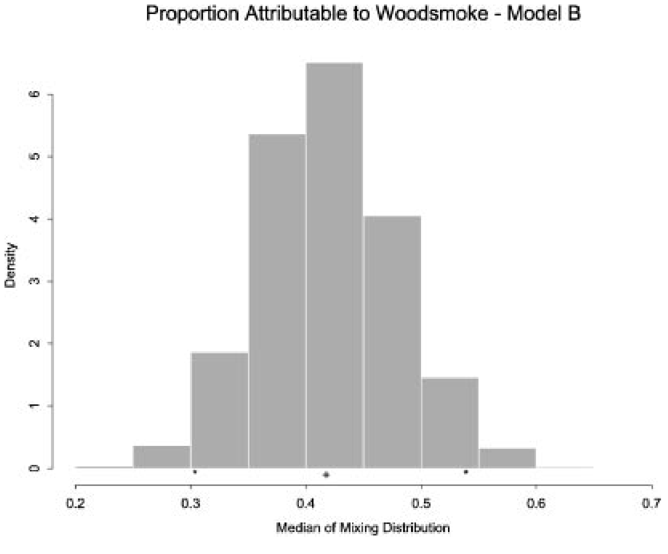

Figure 5. Histogram for median mixing proportions – model B. This histogram shows the distribution of MCMC realizations for the median of the mixing proportion distribution, μ_α. The median of this distribution is approximately 0.42, and a 95% credible interval is (0.30, 0.54)

0.54). This credible interval is roughly 30 per cent narrower than the credible interval produced by model A. As before, allowing increased variability in the source profiles results in reduced variability in the distribution of the mixing parameters.

5. DISCUSSION

In this article, I explore a restricted version of the receptor modeling problem where the number of sources is known, and partial information is available about the individual source profiles. Further, I assume that only relative concentrations of the chemical species are observed. Source profiles and mixing proportions are treated in a hierarchical modeling framework where all quantities, including measurement error, are compositions. Source chemical 'signatures' are outfitted with informative prior distributions summarizing our knowledge of their profiles. Conversely, little is (typically) known of the mixing process. Thus, mixing proportions (and measurement error) are modeled with proper, diffuse prior distributions. Measurement error is incorporated in the model by means of Aitchison's (1986) perturbation operator. Inference is performed using MCMC sampling of the joint posterior distribution.

Among recent statistical contributions to receptor modeling (BR; Park *et al.*, 1999), my approach is most similar to BR. We both address the nuisance parameter problem by modeling mixing proportions as random quantities (Kiefer and Wolfowitz, 1956). Further, we model compositional observations and constrain components of the convex mixture to also be compositional quantities. However, one point of difference is the method by which compositional errors are incorporated in the model. BR uses a Dirichlet convolution process to account for measurement error, while I use a perturbation error structure (explained in detail in Billheimer *et al.*, 1997, 2000). The perturbation construction allows the usual notions of additive error and projection (estimation) to be extended to compositional data in a natural way. Further, it avoids the multidimensional numerical integration required for the Dirichlet convolution model.

By way of contrast, Park *et al.* (1999) model absolute concentrations of chemical species. They assume source profiles to be compositions, but contributions of individual sources and measurement error are not constrained. Further, source contributions (incidental parameters) are treated as fixed, unknown parameters. Park *et al.* (1999) achieve consistent sequence estimators by adapting the quasi-random functional model or replicated functional model of Gleser (1983) to source contribution estimation. However, the resulting estimators for source profiles do not satisfy the summation constraint for compositions, nor are source contribution estimators required to be non-negative.

A second difference concerns assumptions required for model identifiability. Park *et al.* (1999) specify identifiability conditions that rely on 'tracer' species in the source profiles. That is, at least $p - 1$ distinct chemical species must be absent from each source profile. BR identifies model parameters by determining facets of the source polytope. This is accomplished by specifying a subset of observations that lie on facets of the polytope. Here, at least one source does not contribute to each observation on a facet. Parameter estimation is conducted conditionally on the facet observations. In both cases, these assumptions allow model identifiability by reducing the dimensionality of the parameter space.

My approach differs from both BR and Park *et al.* (1999) by using (partial) knowledge about source profiles to define their prior distributions. This is similar to the approach described in Press and Shigemasu (1989) for achieving unique estimates in factor analysis. Informative prior distributions for the source profiles help to 'focus' the posterior distribution, and thus allow parameter estimation. That is, they provide sufficient structure to the posterior distribution to define a maximum over this surface. This approach is quite different from mathematical 'identification' of the model by restricting the parameter space.

Choice of the 'identification method' clearly depends on the problem under consideration and the information avaliable for modeling. In spite of their (seemingly) different approaches to identifiability, these methods are not mutually exclusive. Indeed, it appears that one could combine both methods in a Bayesian framework by specifying prior distributions on source profiles (to reflect tracer elements), or on mixing proportions (to reflect observations on the boundary of the source polytope). Exploration of such combinations of prior information is planned for future work.

Finally, I address issues related to reduced variability in mixing proportion estimation associated with source profile prior distribution(s). The analysis in Section 4 presents a comparison between the Bayesian hierarchical modeling method of this article and the maximum likelihood approach of BR. An important difference in results is a 95 per cent credible interval for the wood-smoke contribution that is approximately 30 per cent shorter than BR's 95 per cent confidence interval. While these intervals are not strictly comparable, they each describe a range of values in which one would expect the parameter. This difference in interval length is likely due to (i) use of an informative prior distribution in the Bayesian model, and (ii) use of an explicit 'additive' error structure, instead of BR's Dirichlet convolution.

The results from model B show that estimation of multiple unknown source profiles is possible in the compositional modeling framework. Allowing increased uncertainty in both source profiles further reduces variability in the mixing parameter's posterior distributions. Such a result is not surprising since increased flexibility in specifying source compositions allows mixing proportions to be more homogeneous in describing observed variability. In model A, where one source profile (motor vehicle emissions) is unknown, the range of plausible values of the proportion attributable to wood-smoke is roughly 30 per cent narrower than that estimated by BR (where the motor vehicle profile was estimated by the MLE). Further, by allowing both source profiles to be unknown, the range of values is reduced by an additional 30 per cent. These results suggest that we need further clarification of the relationship between fixing observations on facets of the source polytope, fixing vertices of the polytope, and specification of a prior distribution for source profiles.

In addition to these issues, the analysis of the Juneau receptor data identifies directions for new methodological developments. First, methods for incorporating and assessing the effect of covariates are needed. Both environmental and anthropogenic factors are likely to be important determinants of source contributions. Quantitative methods for evaluating such covariates would provide powerful tools for understanding observed variation.

Second, most methods used in receptor modeling assume that observations are mutually independent. (Indeed, only Park *et al.* (2000) have attempted to account for serial dependence.) As with other atmospheric data, one expects temporal dependence between multiple observations from a single site, and spatial dependence for a network of samplers. While correlation complicates evaluation of inherent variability, it can be used to benefit prediction. Indeed, one might anticipate substantially improved estimation of the mixing process by 'borrowing strength' from neighboring observations.

Finally, there are many challenges of receptor modeling not directly suggested by the Juneau data analysis. These include treatment of 'below-detection-limit' values (in chemical applications), formal diagnostic procedures for identifying missing sources, and multiple levels of measurement error. We need to accommodate such 'real world' complexities in statistical models for air pollution data.

A1. APPENDIX: LOGISTIC NORMAL DISTRIBUTION AND COMPOSITIONAL ALGEBRA

The Appendix follows the development in Billheimer *et al.* (2000). Compositional data are vectors of proportions describing the relative contributions of each of k categories to the whole. Mathematically, $z = (z_1, z_2, \ldots, z_k)$', where $z_i > 0$, for all $i = 1, 2, \ldots, k$ and $\Sigma_{i=1}^{k} z_i = 1$. Hence, $\mathbf{z}$ is an element of the $(k-1)$-dimensional simplex (∇^{k-1}). Aitchison (1986) introduces the logistic normal (LN) distribution as a framework for analysis of compositional data. These methods rely on the additive logratio transform, $\phi(.)$, to take observations from ∇^{k-1} to $(k-1)$-dimensional Euclidean space ($\Re^{k-1}$). The additive logratio transform of $\mathbf{z} \in \nabla^{k-1}$ to $\Re^{k-1}$ is defined as

$$\phi(\mathbf{z}) = \left[\log\left(\frac{z_1}{z_k}\right), \log\left(\frac{z_2}{z_k}\right), \ldots, \log\left(\frac{z_{k-1}}{z_k}\right)\right]'$$

This transformation is a bijection with inverse transformation denoted by ϕ^{-1}. Aitchison (1986) terms the inverse transformation the additive logistic transform.

Aitchison models the transformed data via the $k-1$ multivariate normal distribution. Assuming multivariate normality of the transformed data induces a distribution on ∇^{k-1}: the logistic normal (LN) distribution. A key benefit of the multivariate normal assumption is that its rich covariance structure transfers to the logistic normal. This allows positive or negative covariances between pairs of the k elements of the composition. In addition, inference tools developed for multivariate normal data can be applied to the transformed compositions.

The LN density function is

$$f(\mathbf{z} \mid \boldsymbol{\mu}, \Sigma) = \left(\frac{1}{2\pi}\right)^{\frac{k-1}{2}} \mid \Sigma \mid^{-\frac{1}{2}} \left(\frac{1}{\prod_{i=1}^{k} z_i}\right) \exp\left[-\frac{1}{2}(\boldsymbol{\theta} - \boldsymbol{\mu})'\Sigma^{-1}\boldsymbol{\theta} - \boldsymbol{\mu})\right]$$

where $\boldsymbol{\theta} = \phi(\mathbf{z})$ for $\mathbf{z} \in \nabla^{k-1}$. We denote the density function by $L^{k-1}(\boldsymbol{\mu}, \Sigma)$. While the parameters depend on the ordering of the elements of $\mathbf{z}$, the density is invariant with respect to permutations of the elements. Aitchison (1986) also establishes moments and other properties of this distribution, including its role as a limit distribution.

Associated with the additive log-ratio transform is a perturbation operator for compositional data (Aitchison, 1986). Perturbations allow an error structure on ∇^{k-1} analogous to the usual additive error model used in other areas of statistics. An observed proportion vector, $\mathbf{z}$, can be modeled as a location vector ($\boldsymbol{\xi}$) 'perturbed' by an error (α). For $\boldsymbol{\xi}, \boldsymbol{\alpha} \in \nabla^{k-1}$,

$$\mathbf{z} = \boldsymbol{\xi} \oplus \boldsymbol{\alpha} = \left(\frac{\xi_1\alpha_1}{\Sigma_{i=1}^{k}\xi_i\alpha_i}, \frac{\xi_2\alpha_2}{\Sigma_{i=1}^{k}\xi_i\alpha_i}, \ldots, \frac{\xi_k\alpha_k}{\Sigma_{i=1}^{k}\xi_i\alpha_i}\right)$$

and $\mathbf{z} \in \nabla^{k-1}$. The vector $\boldsymbol{\alpha}$ need not be an element of ∇^{k-1} for the perturbation operator to be defined. It is sufficient that $\alpha_i > 0$ for all $i = 1, 2, \ldots, k$. Aitchison (1986) shows a number of properties of the perturbation operator including an inverse perturbation, an identity element

$$\mathcal{I}_{k-1} = \left(\frac{1}{k}, \frac{1}{k}, \ldots, \frac{1}{k}\right)'$$

and a power transformation for compositions (Aitchison, 1986, p. 120).

A1.1. *Algebra for compositions*

Following Aitchison's (1986, 1992) suggestion, we use the perturbation operator to define an addition operator for compositions. Further, the power transformation allows us to define scalar multiplication of a composition $\mathbf{z}$ by a scalar a as

$$\mathbf{z}^a = \left(\frac{z_1^a}{\Sigma_{i=1}^k z_i^a}, \frac{z_2^a}{\Sigma_{i=1}^k z_i^a}, \ldots, \frac{z_k^a}{\Sigma_{i=1}^k z_i^a} \right)$$

Billheimer *et al.* (2000) show that ∇^{k-1} equipped with the perturbation operator and scalar multiplication constitutes a complete inner product space. This additional mathematical abstraction allows the definition of a norm on ∇^{k-1}. Further, it provides a framework for algebraic operations on compositions. The inner product and norm are defined as follows.

Definition A. 1. *For* $\mathbf{u}, \mathbf{z} \in \nabla^{k-1}$, *let* $\boldsymbol{\theta} = \phi(\mathbf{u})$ *and* $\boldsymbol{\eta} = \phi(\mathbf{z})$. *Define by*

$$\langle \mathbf{u}, \mathbf{z} \rangle = \boldsymbol{\theta}' \boldsymbol{\mathcal{N}}^{-1} \boldsymbol{\eta}$$

the inner product of $\mathbf{u}$ and $\mathbf{z}$.

Here, $\mathcal{N} = [I_{k-1} + \boldsymbol{j}_{k-1}\boldsymbol{j}'_{k-1}]$, where I_{k-1} is a $(k-1)$-dimensional identity matrix, and $\boldsymbol{j}_{k-1}$ is a $k-1$ column vector of ones. Note that

$$\mathcal{N}^{-1} = I_{k-1} - \frac{1}{k} \boldsymbol{j}_{k-1}\boldsymbol{j}'_{k-1}$$

Definition A. 2. *Define the norm for* $\mathbf{u} \in \nabla^{k-1}$, $\| \mathbf{u} \|$, *by* $\langle \mathbf{u}, \mathbf{u} \rangle^{1/2}$.
Inclusion of the matrix $\mathcal{N}^{-1}$ ensures that the inner product and norm are invariant to permutations of elements of $\mathbf{u}$. Note also that the norm defined above is a sum of squares of log-ratios. This definition is contained in the class of functions meeting Aitchison's (1992) criteria for a compositional metric.

A.1.2. *Differences between compositions*

The definition of an (inverse) addition operation and a norm allow us to measure the difference between compositions. For demonstration, consider three compositions in $\nabla^2, \mathbf{z}_1 = \mathcal{I}_2 = (1/3,\ 1/3, 1/3), \mathbf{z}_2 = (0.80, 0.10, 0.10)$, and $\mathbf{z}_3 = (0.98, 0.01, 0.01)$,

We first note that the norms of these compositions are

$$\| \mathbf{z}_1 \| = 0, \quad \| \mathbf{z}_2 \| = 1.698, \quad \text{and} \quad \| \mathbf{z}_3 \| = 3.744$$

Thus, the defined norm measures the distance of a composition from $\mathcal{I}_{k-1}$, the 'center' of ∇^{k-1}.

Next, using the inverse of the perturbation operator, we find the difference between pairs $\mathbf{z}_1$ and $\mathbf{z}_2$, and $\mathbf{z}_2$ and $\mathbf{z}_3$. To find the difference between two compositions we perturb the second by the elementwise inverse of the first. That is,

$$\mathbf{z}_2 \ominus \mathbf{z}_1 = \mathbf{z}_2 \oplus \mathbf{z}_1^{-1} = \mathbf{z}_2$$

since $\mathbf{z}_1$ is the identity element. Similarly,

$$\mathbf{z}_3 \ominus \mathbf{z}_2 = \left(\frac{[\mathbf{z}_3]_1 [\mathbf{z}_2]_1^{-1}}{\sum_{i=1}^3 [\mathbf{z}_3]_i [\mathbf{z}_2]_i^{-1}}, \frac{[\mathbf{z}_3]_2 [\mathbf{z}_2]_2^{-1}}{\sum_{i=1}^3 [\mathbf{z}_3]_i [\mathbf{z}_2]_i^{-1}}, \frac{[\mathbf{z}_3]_3 [\mathbf{z}_2]_3^{-1}}{\sum_{i=1}^3 [\mathbf{z}_3]_i [\mathbf{z}_2]_i^{-1}} \right)$$
$$= (0.860, 0.070, 0.070)$$

where $[\mathbf{z}_i]_j$ is the *jth* element of the compositon $\mathbf{z}_i$. Thus, (0.86, 0.07, 0.07) is the composition by which we need to perturb $\mathbf{z}_2$ to obtain $\mathbf{z}_3$. By taking the norm of the difference composition, we measure the distance between $\mathbf{z}_2$ and $\mathbf{z}_3$,

$$\|\mathbf{z}_3 \ominus \mathbf{z}_2\| = \|(0.86, 0.07, 0.07)\| = 2.046$$

Note that the distance from $\mathbf{z}_1$to $\mathbf{z}_2$ is 1.698, while the distance from $\mathbf{z}_2$ to $\mathbf{z}_3$ is larger at 2.046. For additional details of the compositional algebra and proofs of its Hilbert space characteristics, see Billheimer *et al.* (2000).

REFERENCES

Aitchison J. 1986. *The Statistical Analysis of Compositional Data.* Chapman & Hall: New York.

Aitchison J. 1992. On criteria for measures of compositional difference. *Mathematical Geology* **24**: 365–379

Aitchison J, Bacon-Shone J. 1999. Convex linear combinations of compositions. *Biometrika* **86**: 351–364.

Aldershof B, Ruppert D. 1987. A statistical analysis of woodstove PAH emissions and source apportionment of ambient air samples. Unpublished EPA report: Research Triangle Park.

Bandeen-Roche K. 1994. Resolution of additive mixtures into source components and contributions: a compositional approach. *Journal of the American Statistical Association* **89**: 1450–1458.

Besag JE. 1974. Spatial interaction and the statistical analysis of lattice systems (with Discussion). *Journal of the Royal Statistical Society B* **36**: 192–236.

Besag JE, Green PJ. 1993. Spatial statistics and Bayesian computation (with Discussion). *Journal of the Royal Statistical Society B*, **55**: 25–37.

Billheimer D, Fagan WF, Guttorp P. 2001. Statistical interpretation of species composition. *Journal of the American Statistical Association.* (at press).

Billheimer D, Cardoso T, Freeman E, Guttorp P, Ko H, Silkey M. 1997. Natural variability of benthic species composition in the Delaware Bay. *Journal of Environmental and Ecological Statististics* **4**: 95–115.

Gleser LJ. 1983. Functional, structural and ultrastructural Errors-in-variables Models. *ASA Proceedings of the Business and Economic Statistics Section.* American Statistical Association: Alexandria, VA: 57–66.

Hopke PK (ed.). 1991. *Receptor Models for Air Quality Management.* Elsevier Science Publishers: Amsterdam.

Hopke PK. 1999. An introduction to source receptor modeling. In *Elemental Analysis of Airborne Particles.* Landsberger S, Creatchman M (eds). Gordon and Breach Science Publishers: Amsterdam.

Kiefer J, Wolfowitz J. 1956. Consistency of the maximum likelihood estimator in the presence of infinitely many incidental parameters. *Annals of Mathematical Statistics* **27**: 887–906.

Kowalczyk GS, Choquette CE, Gordon GE. 1978. Chemical element balances and identification of air pollution sources in Wasington DC. *Atmospheric Environment* **12**: 1143–1153.

Lindsay BG. 1983. The geometry of mixture likelihoods, I: General theory. *Annals of Statistics* **11**: 86–94.

Mardia KV. 1988. Multidimensional multivariate Gaussian Markov random fields with applications to image processing. *Journal of Multivariate Analysis* **24**: 265–284.

Neyman J, Scott EL. 1948. Consistent estimates based on partially consistent observations. *Econometrica* **16**: 1–32.

Park ES, Guttorp P, Henry RC. 2000. Multivariate receptor modeling for temporally correlated data by using MCMC. *Technical Report Series, No. 043. National Research Center for Statistics and the Environment.*

Park ES, Spiegelman CH, Henry RC. 1999. Bilinear estimation of pollution source profiles in receptor models. *Technical Report Series, No. 109. National Research Center for Statistics and the Environment. No. 109.*

Press SJ, Shigemasu K. 1989. Bayesian inference in factor analysis. In *Contributions to Probability and Statistics: Essays in Honor of Ingram Olkin.* Gleser LJ, Perlman MD, Press SJ, Sampson AR (eds.). Springer-Verlag: New York: 271–287.

Raftery AE, Lewis SM. 1992. How many iterations in the Gibbs sampler? *Bayesian Statistics 4, Proceedings of the Fourth Valencia International Meeting.* Clarendon Press: Oxford: 763–773.

Reiersol O. 1950. Identifiability of a linear relationship between variables which are subject to error. *Econometrica* **18**: 375–389.